DOCUMENT RESUME

ED 220 494 TM 820 507

AUTHOR Brown, William J.
TITLE Considerations in Conducting Large-Scale Writing Assessments as Part of Minimum Competency Testing.
PUB DATE [Apr 82]
NOTE 12p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Criteria; Field Studies; Graduation Requirements; High School Students; Holistic Approach; Minimum Competencies; *Minimum Competency Testing; *Program Implementation; Scoring Formulas; State School District Relationship; *State Standards; Test Construction; *Testing Programs; *Test Reliability; *Writing Evaluation; Writing Skills
IDENTIFIERS *North Carolina

ABSTRACT

In 1980-81 field tests were undertaken in North Carolina to develop a procedure for making writing competency a minimum requirement for graduation from high school. Measuring writing skills with adequate validity and reliability requires planning for data collection and varied scoring methods, personnel training, dissemination of assessment results for improved instruction, and the cost of assessment. Several prototype measurement models using ninth and eleventh grade students examined samples of message notes, an expository passage, a business letter and completing an application form. The efficacy of the measurement models relied on consideration of a scoring method to establish reliability with an impressionistic approach labeled "focused holistic" within scoring criteria to be consistently applied across central or regional scoring sites. Writing remediation needs were also investigated. A second year of design reconsiderations will include the effort to increase reliability by discriminating between adjacent score values to establish a cut-off score; the expansion to direct and indirect writing measures; the development of exercise review procedures; and decisions regarding minimum competency objectives. (CM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED220494

CONSIDERATIONS IN CONDUCTING LARGE-SCALE WRITING ASSESSMENTS AS PART OF MINIMUM COMPETENCY TESTING

By William J. Brown
Special Assistant for Research
North Carolina Department of Public Instruction
Raleigh, North Carolina 27611

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

I. Introduction

Writing is often judged as an important basic skill along with reading and mathematics; but because of the difficulty in measuring and evaluating writing, most statewide assessments have not included direct measures of writing. Instead, assessments are made of students' proficiency in language arts as measured by multiple choice tests. Some states have measured writing directly and on a statewide basis. Their efforts have contributed greatly to the increased interest which other states are manifesting now. However, states who want to venture into the measuring of writing are finding that these early efforts in other states are only partially generalizable as prototypic models. Thus, considerable effort must still be made by each state to develop its own procedure for measuring writing that is both credible and feasible.

In North Carolina writing is being strongly considered as an area for general assessment at grades three, six, and nine and also as a minimum competency which must be demonstrated prior to high school graduation. The complexity of the task and the enormous costs have caused a slow but steady pace toward implementation. This paper will describe the early field tests associated with adding writing as a minimum competency for high school graduation. Although the field tests were focused upon minimum competencies, at least one of the approaches considered has sufficient generalizibility so that it can be extrapolated to assessing writing at any level of proficiency.

TM 820 582

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
W. J. Brown
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

If it is determined that writing can be measured with adequate validity and reliability, the minimum competency skills will be expanded to include writing as well as reading and mathematics which are presently being assessed. The responsibility for making a recommendation to add a skill area to the minimum competency program in North Carolina lies with the 15 member Competency Testing Commission which was created by the Legislature and appointed by the Governor. The Commission's recommendations go to the State Board of Education where final action is taken.

II. Problems in Implementing a Writing Assessment

Implementing a writing assessment, even on a small scale, has numerous problems which are associated with validity, reliability, and reaching a concensus on the writing skills to be measured. These problems are magnified when considering a large scale assessment and are expanded to include considerations related to collecting and scoring the writing samples, to securing and training personnel in scoring procedures, and to disseminating and using the results from the writing assessment in improving instruction. The high costs of conducting a writing assessment tops off the lists of problems and concerns.

Given these problems, the department has completed a year of field testing several prototypic models and has initiated a second year of field testing unresolved problems. During the first year of field testing, writing samples were taken from approximately 1,000 eleventh grade students, 180 exceptional children at grade eleven, and 180 ninth graders.

The 1980-81 field test examined the efficacy of measurement models which included up to four writing samples and two scorings for each student. The reliabilities of every combination of four samples and two scorings were compared against the complete model. In addition, models were examined which included screening devices to lower the number of complete scorings required.

The field test also examined corollary questions such as:

(1) the effect of scoring the papers centrally by teachers from throughout the state as contrasted to scoring the papers regionally using the teachers from the schools of students in the sample,

(2) the intra-judge reliability at the central scoring site,

(3) the inter-scorer reliability at the central site where scorers were randomly assigned their papers,

(4) the correlations of total item score to the external criterion scores of estimated performance, English grades, and reading scores from the competency test,

(5) the distributions of item scores and total scores for regular eleventh graders, exceptional eleventh graders, and regular ninth graders, and

(6) the test-re-test correlations.

In the 1980-81 field test, students were measured on four types of writing:

(1) their ability to listen to a message, take notes, and then write a note giving the details of the message,

(2) their ability to write an expository paragraph,

(3) their ability to write a business letter, and

(4) their ability to complete an application form.

The scoring methods included an impressionistic approach which has been labeled focused holistic. This scoring procedure yielded an overall score of 4, 3, 2, 1 or 0. These scores reflected an overall impression of the quality of the paper after focusing upon specified criteria. This variation from holistic scoring was done so that criteria for score

values would be consistent across scoring sites and times. The Commission felt that scoring criteria must be specified and consistently measured across scoring sites since one of the values would ultimately be chosen as a cut-off score for passing.

An analytic scoring procedure was also tested because of the need to provide remediation for students who failed the test. The score values for each criteria were to be reported in a manner that would allow for instruction in areas of punctuation, capitalization, spelling, main idea, supporting detail and the like.

The field test also was to generate cost estimates which would allow budget requests to be made. The estimates came from measuring the time required for training scorers, for the actual scoring of papers, and for the clerical tasks associated with the scoring process.

It was anticipated that a second field test in 1981-82 would be necessary to refine the most feasible models identified in the first field test.

III. Results from the 1980-81 Field Test

The 1980-81 field test results provided a wealth of information on the measurement of writing. The results are described fully in the Contractor's report to the Commission entitled, "North Carolina Fall 1980-81 Writing Field Test - Final Report." From this report, the following conclusions were drawn:

1. There was considerable variability among the four writing measures - the note, the paragraph, the business letter, and the application form. The note and the application form yielded the highest mean scores. Of the four, the business letter was most difficult. The scores of eleventh graders and ninth graders were comparable. The exceptional children did poorly on the note, the paragraph and the

business letter. Their highest achievement was on the application form.

2. The reliability estimates of a student taking the writing test twice were produced from a series of generalizability studies which combined various combinations of the four items and the two scorers. The reliability increased as more items were added and as more scorers were added. Apparently, more reliability was added by increasing the number of items than in increasing the number of times the items were scored. The optimal point of combination appeared to reside in three or four items and two scorings. Even so, the reliabilities varied considerably depending upon the combination of items which were being considered. When three items and two scorings were evaluated, reliabilities of .740 were obtained from the paragraph, the business letter, and the application form and from the note, the paragraph, and the business letter.

   When the item types were restricted to just two types, the best combinations were taking two samples of a paragraph in combination with an application form (.819) and two samples of a note in combination with a paragraph (.830).

3. Although a cut-off score did not exist at the time of the field test, the Contractor was instructed to select three arbitrary points between 1.5 and 2.5 as cut-offs so that the screening procedures could be evaluated. Six screening models were used but each one had significant numbers of false positives (up to 17 percent). These are students who pass the screening device when they would have failed if the full model had been applied. The number of false negatives (up to 11 percent) were lower. The best results were

obtained from screening models that included more measurement and would thereby be higher in cost and less of an advantage over the full model.

4. The correlations between the scores of central scorers and regional scorers showed a relatively high degree of association for the total test (.932) and for the individual items.
5. The intra-judge reliability ranged from .784 for the note to .948 for the business letter. The best pair for reliability was the business letter and the application form.
6. The inter-scoring reliability for the total score was .372 and for individual items ranged from .672 for the note to .767 for the application form.
7. The correlations of total item scores to external criteria were identified as a means of determining validity. These item scores were correlated with informed teacher estimates of writing, English grades, and reading test scores each of which could be inferred to be related at least partly to writing ability. The total writing scores were correlated .575 to teacher's estimates, .415 to English grades, and .623 to reading scores.
8. The test-retest correlations came from retesting four classes of eleventh graders after about three weeks had elapsed. The scorers for the retest were chosen from among the central scorers but no attempt was made to match the initial scorer and the retest scorer. The correlation resulting from this analysis was .24 for the note, .74 for the paragraph, .54 for the business letter, .35 for the application form, and .48 for the total score.

After reviewing these results, the Commission solicited the advice of persons who had technical expertise in measurement and statistics. After several sessions of review and considering alternatives, the Commission concluded that sufficient reliability had been demonstrated if the results were to be used as group information for curriculum review or planning. It was concluded, however, that more reliable procedures must be developed before consequential decisions could be made about individuals. The Commission further concluded that the reliability of a writing score at the cut-off score was the important factor for writing scores which were to be used in minimum competency programs. To that end, the Competency Commission has refocused its attention to determining what is minimum competence in writing and how to discriminate between those students who have that competence and those who do not.

IV. Implications for Further Field Testing

Although the results from the first field study indicated that writing could be measured with a moderately high reliability, there was not sufficient reliability to discriminate between adjacent score values. For minimum competency programs, it is imperative that judgments near the passing score are reliable measurements. From the field study data, it was determined that scores of 1 were not likely to be scores of 3 or 4, but it was probable that many of them would have been scored a 2 if they were re-scored or if the student re-took the test without instruction. After continued study of the data, it was concluded that additional refinements to the scoring procedure would add some reliability but not enough to overcome the overlap between adjacent scores. Thus, the Commission restudied the scoring design and considered alternative measurement approaches.

After a reanalysis of the holistic and analytic scores, it was concluded that the note and the expository paragraph were similar writing tasks as were the application and the business letter. The Commission also re-examined the objectives which it considered as minimum and sought alternative measurement approaches. The following considerations gave direction to the design of the second field study:

1. The scoring process could be made more reliable at the cut-off score if the scoring procedure focused upon identifying papers which were below minimum competence and those which were above that level. No distinctions were to be made on how poorly or how well a paper was written.
2. The measurement of writing was to be expanded to include both direct and indirect measures of writing. The direct measure of writing was to focus upon writing samples which were scored principally for writing thought and communication. A multiple choice test was to indirectly measure writing mechanics and usage.
3. The decision-making process about minimum competence in writing would combine both measures of writing into a score which was sufficiently reliable for individual decision making. It was also noted that at least four opportunities are provided for testing so that testing error alone would not deny a student from earning a diploma, even with some test unreliability.
4. The writing exercises would be developed so that a student always had a choice of material, thereby improving student motivation. The exercises were to be reviewed thoroughly to insure that they were within the experience and knowledge range of tenth or eleventh grade students.

The Commission conducted work sessions in preparation for letting a request for proposals. These sessions developed two surveys of teachers which were to ascertain the relative importance placed upon the four writing types (notes, paragraphs, letters, forms), the importance of objectives related to each piece of writing, and a general survey of writing status and practice. Information from these surveys would add to the concensus being developed as to what is important and they would assist the Commission and the Department in planning for a statewide implementation of writing.

The next steps of the Commission were to analyze the writing samples from the first field study in relation to the minimum competence expected for high school graduates. To accomplish this, the Commission scored approximately 192 items of regular eleventh graders and 55 items of exceptional students. The items were scored as below or above minimum competence. When there was any disagreement among the four scorers, the item was scored again independently without knowledge of the previous scoring. If any disagreement continued, the item was discussed and a jury rating was assigned to it. The final rating of pass or fail was recorded and the two sets of papers became criterion papers against which scoring criteria could be developed during the field study.

During this phase of the Commission's work, it became apparent that most papers which had earlier been scored as either a 3 or a 4 were judged as passing papers by the Commission without a second reading. Eighty-three percent of these papers were passed without a second reading. Of the 19 papers without initial agreement, 11 were resolved by a second reading and 9 were resolved by concensus. In the final analysis, only one paper from this group failed.

There was more disagreement on the papers which had been scored as twos. Of the 50 papers in this group, there was initial disagreement among the four scorers on 36 papers. Thirteen of these disagreements were settled by a second reading and 23 resolved by a third reading. Thirty-three of the 50 papers ultimately passed.

There was disagreement on about four out of ten papers that had been scored a one. There was initial disagreement on 22 of the 54 papers in this group. Thirteen of the 22 disagreements were settled on the second reading with nine requiring final resolution. Nine of the 54 papers in this group passed.

The rank order of the writing types in terms of low disagreement was as follows: note (33.3%), business letter (33.3%), expository paragraph (45.8%), and application form (47.9%).

The conclusions drawn thus far are that good writers are identified easily and reliably. Marginal writers have major faults in their writing thought and/or writing mechanics, but only half of them write so poorly as to be judged incompetent. Likewise, some poor writers get their message across clearly enough to be judged as competent.

Hopefully, the addition of a multiple choice writing test component will help in discriminating among these marginal writers. If not, the key decisions to be made are:

1. Should the standards in writing be high so that writing competence is assured for those above the standard, both in terms of validity and reliability?
2. Should a moderately low standard be set as a competency requirement, and set a standard which identifies additional students as just passing the minimum requirement? These students would also be

given the opportunity for remediation but would not face the writing competency test again.

3. Should writing be pursued as an assessment but not as a competency requirement?

These considerations can be pondered as plans are made for the final field study.